Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset

Shou-de Lin
Computer Science Department
University of Southern California
sdlin@isi.edu

Hans Chalupsky
Information Sciences Institute
University of Southern California
hans@isi.edu

ABSTRACT

This paper describes a submission to the Open Task of the 2003 KDD Cup. For this task contestants were asked to devise their own questions about the HEP-Th bibliography dataset, and the most interesting result would be selected as the winner. Instead of taking a more traditional approach such as starting with a inspection of the data, formulating questions or hypotheses interesting to us and then devising an analysis and approach to answer these questions, we tried to go a different route: can we develop a program that automatically finds interesting facts and connections in the data?

To do this we developed a set of unsupervised link discovery methods that compute interestingness based on a notion of "rarity" and "abnormality". The experiments performed on the HEP-Th dataset show that our approaches are able to automatically uncover interesting hidden connections (e.g. significant relationships between people) and unexpected facts (e.g. citation loops) without the support of any prerequisite knowledge or training examples. The interestingness of some of our results is self-evident. For others we were able to verify them by looking for supporting evidence on the World-Wide-Web, which shows that our methods can find connections between entities that actually are interestingly connected in the real world in an unsupervised way.

1. INTRODUCTION

In the Open Task of the KDDCup 2003 contestants were asked to devise their own questions about the High Energy Physics-Theory (HEP-Th) bibliography dataset, and the most interesting result would be selected as the winner. A traditional approach to such a problem would be to start with a manual inspection of the dataset, formulate some interesting hypotheses or questions based on the observations, and then devise an analysis to verify the hypotheses or answer the questions.

For our submission, we tried to turn this traditional methodology around and asked: "instead of manually specifying and addressing some interesting problem, can we have an unsupervised program that finds interesting facts and connections in the dataset automatically?" This is of course also a question about the dataset, but it is at a more general, abstract level. To achieve this we developed various domain independent unsupervised link discovery methods that can detect interesting facts and connections automatically. Once such interesting instances or connections have been discovered, we can inspect them manually to understand why the program selected them. This process will then sometimes inspire the hypothesis of patterns or rules underlying the dataset, which could then be verified by some other KDD methods.

Using Rule-based and supervised learning approaches to mine interesting instances or events are limited by some lack of objectivity and generality. For a rule-based system, the rules or patterns for "interestingness" contain subjective biases of the people who generated them. For a supervised learning system the selection of training data embodies a strong bias on what is to be considered interesting in the domain. Additionally, both approaches lead to highly domain-dependent solutions that need to be adapted whenever the domain changes (e.g., via adapting rules or training on new examples relevant to the new domain).

To overcome these limitations, we propose a set of **unsupervised** link discovery methods to detect interesting facts in a dataset. We utilize a notion of "**rarity**" to measure the interestingness of paths in the data and define four different rarity measures to accommodate different points of view. Path rarity forms the basis for "loop rarity" when we look for interesting circular paths in a dataset. Finally, we look for "**abnormal**" combinations of paths based on their rarity values to detect interesting nodes connected to a source node. Using this approach we can answer the following types of queries:

1. Which nodes (e.g., people, organizations, journals, papers, keywords, etc.) are interestingly connected to a given node?

2. Which path or loop between two nodes or entities is an interesting one from various points of view?

2. UNSUPERVISED LINK DISCOVERY METHODS

2.1 Definitions and Assumptions

We focus on discovering interesting facts from datasets that can be represented as sets of entities connected by a set of binary relations. In other words, each object in the dataset is treated as a separate entity and there are different types of binary relations connecting these entities. This kind of data can naturally be represented by a labeled graph as the one shown in Figure 1, where nodes stand for entities and links for binary relations. For example, we can represent a bibliography dataset this way by modeling papers, authors, organizations, etc. as nodes and authorship, citation, etc. as relations. We also assume that the data employs a fairly rich vocabulary of relations where different link types represent different semantic relationships.

2.2 Novel Path Discovery via Rarity Analysis

Under these assumptions, we define the novel path discovery [1] problem as follows: given an arbitrary pair of entities in a network with numerous paths connecting them, find interesting paths between them. One challenge of this problem is that the interestingness of a path is non-linearly related to the interestingness of its individual links. That is, each individual link of a path might not be interesting at all but it is the combination of them that represents something special.

To deal with novel path discovery problems, we observe that to some extent **rarity** is an indicator of **interestingness.** That is, an event that occurs infrequently compared to other events has the potential to be interesting and thus worth being reported. Using rarity as a measure for interestingness fulfills the need of capturing domain specificity: the same event can be rare in one

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
2003

2. REPORT TYPE

3. DATES COVERED
00-00-2003 to 00-00-2003

4. TITLE AND SUBTITLE
Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California,Information Sciences Institute ,4676 Admiralty Way,Marina del Rey,CA,90292

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
unclassified	**unclassified**	**unclassified**		**6**	

domain but not in the other. For example, the event "A cites B's paper" could be more interesting in a criminal database than "A kills B" because it occurs rarely, despite the fact that people might generally think that killing is the more interesting event. Rarity is also flexible enough to handle different points of view. For example "A cites B's paper" can be rare from A's point of view but not from B's point of view due to the fact that A rarely cites others but B is commonly cited by many other people.

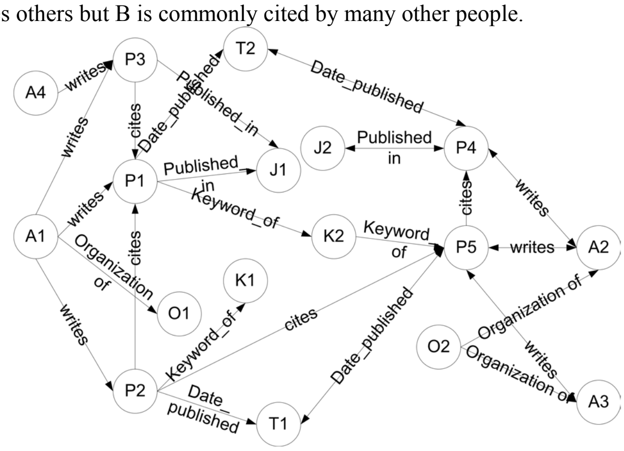

Figure 1: A bibliography network with 16 nodes and 21 links

To apply these ideas to the novel path discovery problem, we have to define rarity measurements for paths in the network. Note that in a multi-relational network as shown in Figure 1, every path occurs exactly once, thus all of them are equally rare. We therefore need a more relaxed definition to measure path rarity. We do this by defining the rarity of a path as the reciprocal of the number of **similar** paths to it. We accommodate view dependency by defining four different measures based on different views of similarity.

An *n*-step path can be defined as a combination of *n+1* entities and *n* relations $e_0 \xrightarrow{r_0} e_1 \xrightarrow{r_1} e_2 .... \xrightarrow{r_{n-1}} e_n$ where ei's are entities (or nodes) and r_i's are relations (or links). We call e_0 the source and e_n the target of the path. Note that in the novel path discovery problem we do not consider paths that contain one or more loops. We define the **type** of a path t_p as the ordered sequences of relations [$r_0....r_{n-1}$] of the path *p*. For example, the path "A writes a paper that cites a paper published at time T1" and the path "B writes a paper that cites a paper published at time T2" are of the same type [writes, cites, date_published].

Path rarity of a path *p* is defined as ***rarity$_i$(p) = 1/N$_i$(p), 1≤ i ≤ 4*** where $N_i(p)$ (or simply N_i) is the number of paths similar to *p* according to a particular view on path similarity.

We define N1 as the number of paths that have the same type as *p* as well as the same source node e_0 and target node e_n. According to this view, the rarity of the path "A1 is the author of P2 and P2 cites P1" in Figure 1 is 1/2, since there is only one other path "A1 is the author of P3 and P3 cites P1" that is similar to it.

We define N2 as the number of paths that have the same type as *p* as well as the same source node e_0. According to this view, the rarity of the example path described above is 1/3, since there is one more path "A1 is the author of P2 and P2 cites P5" that matches the criteria.

Similar to N2, N3 is the number of paths that have the same type as *p* as well as the same target node e_n. Using this measure, the example path has rarity 1/3, since besides the paths that satisfy N1 rarity, there is one more path "A4 is the author of P3 and P3 cites P1" that matches the criteria.

Finally, we define N4 as the number of paths that simply have the same type as *p*. Under this view the rarity of the example path is 1/5, since there are five paths in Figure 1 of type "X is the author of Y and Y cites Z".

With these rarity measures in hand, we have a systematic way to answer a query such as "what is the most interesting path between nodes X and Y?" We simply enumerate all paths between X and Y and return the one with the highest rarity value. By using rarity to determine the most interesting path, we not only take domain specificity and user views into consideration, but also avoid being misled by the apparent meaning of the links.

2.3 Novel Loop Discovery

The novel loop discovery problem is a variation of novel path discovery. It aims at finding interesting loop paths such as the following:

$$e_0 \xrightarrow{r_0} e_1 \xrightarrow{r_1} e_2 .... \xrightarrow{r_{n-1}} e_0$$

Loop rarity is measured similar to path rarity. However, since in a loop the source is identical to the target, the N1, N2, and N3 value will all be the same. Thus, there are only two different loop rarity measurements: 1/N1 measures how rare a specific loop is from e_0's point of view and 1/N4 determines how rare this type of loop is in general.

2.4 Novel Node Discovery

The goal of novel node discovery is to find entities that are interestingly connected to a given source entity. To detect such interesting nodes, we generalize the concept of "rarity" to "abnormality". We claim that two nodes are interestingly connected to each other if there are **abnormal** connections between them. Abnormality is a relative measure and in this paper we defined it based on the **contribution** of different path types from other notes to the source.

Definition: The **contribution** of a path type t_p for some path *p* going from an arbitrary node *T* to the source node S is the number of paths with the same type t_p connecting the source and the target (or the N1 value) divided by the number of paths with the same type emanating from the source and leading to an arbitrary target (or the N2 value).

The contribution N1/N2, which is between 0 and 1, is in fact the conditional probability that represents "if one picks a path with path type t_p that emanates from a source *S*, what is the chance this path will terminate at a target *T*". Since this represents how often a path of type t_p emanating from *S* winds up at *T* as opposed to any other node, we also call this value *T*'s contribution to *S* relative to t_p.

Let us motivate why the abnormal contributions might indicate interesting source/target relationships. Assume there is a small dataset of 12 people with just one type of path "co-authorship", and the person S co-authored with all the other 11 people. If S writes 10 papers with each of them except one person T, whom he or she writes 100 papers with (that is, the contribution of T to S with respect to this path type is 0.5 while the other ten people contribute only 0.05 to S), then the relationship of T to S is different from the other ten and therefore more interesting. The same would be the case if S wrote only one paper with T compared to the ten written with each of the other ones.

To be more general, it is necessary to consider abnormal "combinations of paths" instead of just one single type of path. For example if two types of paths are positively correlated (that is if the contribution of one is high, the contribution of the other is also high for the majority of nodes connected to source S, and vice versa), then a node T would be interestingly connected to S if it has high contribution in one type but low in the other. In general each type of path represents some form of "behavior" and the N1/N2 value of that path type signifies how much the target T contributes to source S with respect to this certain behavior. Thus, intuitively, a node could be interestingly connected to S if it contributes differently to S compared with other nodes for various behaviors. In the following, we propose a general unsupervised method to discover novel nodes in a multi-relational dataset by finding outliers in a domain that employs path types as features and their contribution as feature values. The method has five steps:

1. For a target node T in the network, enumerate all the path-types t_p between it and the source node S.
2. For each type of path t_p between S and a target T, compute the contribution N1/N2.
3. View each path type as a feature of T and use its contribution as the feature value. For the path types that do not exist between S and T, assign 0 as the feature value (if a path type does not exist between S and T, the conditional probability is 0).
4. Repeat steps 1-3 on all possible target nodes T in the network.
5. Assume there are a total of m nodes in the dataset and n different types of path in the network, therefore after step 4 we have m-1 n-dimensional points. Then a distance-based outlier-detection algorithm is chosen (in the experiment we used Ramaswamy's k-th nearest distance-based algorithm [2]) to discover the outliers among these m-1 points. Ramaswamy's algorithm ranks the outlier points by their Euclidean distance to the k-th nearest neighborhood. That is, the outliers are those far away from their k-th neighbors (it is allowable to have k points around an outlier point).

Our outlier detector now discovers nodes that have abnormal contribution to the source node in the contribution domain that we consider to be potentially interesting. Through this method we are able to transform this non-numerical, multi-relational mining problem into a typical outlier-detection problem by using path types as features and their contribution as feature values.

3. EXPERIMENTS

Below we describe a set of experiments on the "High Energy Physics - Theory" (HEP-Th) bibliographic dataset to illustrate the validity and usefulness of our novel link discovery methods.

3.1 Data Modeling and Information Extraction

We extracted six different types of nodes and six types of links from the dataset. Nodes represent paper IDs (29014), author names (12755), journal names (267), organization names (963), keywords (40) and the publication time encoded as year/season pairs (60). Numbers in parentheses indicate the number of different entities for each type in the dataset. We defined the following types of links to connect nodes:

writes(a, p) : connects author a to one of his/her papers p.
date_published(p, d) : connects paper p to its publication date d.
organization_of(a, o) : connects author a to an organization o they belong to.
published_in(p, j) : connects paper p to journal j it appears in.
cites(p, r): connects paper p to a paper r it cites.
keyword_of(p,k) : connects paper p to keyword k in its title.

All of these links are viewed to be directional with an implicit inverse link. Thus, there are essentially 12 different relations in the dataset.

We extracted the information from three different types of files; the dates are extracted from the SLAC-date file; the citation information is from the HEP-Th citation file and the rest of the information is extracted from the HEP-Th abstract files. It is straightforward to determine the paper IDs, author names and journal names from the abstract files, since the relevant information has been explicitly annotated. Different spellings of the names were not consolidated and resulted in multiple nodes. The organization of a person is extracted from the contacting email address. That is, a person is viewed to belong to an organization if s/he has ever submitted a HEP-Th paper through an email address that belongs to that organization. A person could belong to multiple organizations if he/she used multiple email addresses. For keywords, we first calculated 40 bigrams (two consecutive words) with the highest probability from all the paper titles (discarding stop words), and then checked the existence of these keywords for each paper title.

The network generated is similar to the one in Figure 1, only that there are 43095 different nodes and 477423 links overall. We then applied the three analyses described in Section 2. Because of space limitations, we only describe the novel node and novel loop discovery results in this report.

3.2 Novel Node Discovery Results

In novel node discovery, we try to find nodes interestingly connected to a source where interestingness is modeled as abnormality. Abnormality is defined based on the contribution N1/N2 of various path types from a target to a source. That is, our novel node discovery program picks nodes that contribute abnormally to the source by detecting outliers in a space defined by the path types and their contribution to the source. In the experiment we limited the path length to be at most four and chose k=1 for the k-th nearest distance-based algorithm.

A convincing evaluation for our novel node discovery system would be to rank the interesting nodes and then evaluate whether the top several nodes carry interesting information. However, evaluating "interestingness" is challenging, since we are faced with a chicken-and-egg problem. To fairly evaluate the interestingness of our results, we would need a set of independent and unbiased criteria to judge whether some discovered fact or connection is actually interesting. But if there were such unbiased criteria, we could simply implement them as our interesting-fact finder.

In this experiment we adopt two orthogonal means to evaluate the discovered results: Once the interesting nodes are determined, we first examine the original network (internal source) to learn the reason why our outlier detector prefers these nodes (note that our system does not have any semantic knowledge). In some cases the interestingness of the results are self-evident semantically (e.g. the examples given in Section 2.4). The second means of evaluation is to use the Web (external source) to find supporting evidence. Since the nodes represent real-world entities such as people, we can "verify" the computed results by investigating whether they reflected real-world, significant connections visible through the World-Wide Web.

We started by picking C.N. Pope as the source node, since in this dataset he is the one with the most publications which provides us with a rich number of connections to other nodes. The first query we wanted to answer was **"which people are interestingly connected to C.N. Pope?"** Our program first enumerates all the path types emanating from Pope (there are 14 of them within four steps). Then for each person node connected to Pope, it computes its contribution to the source for each of these 14 types. Overall there are 12755 people in our dataset, thus, our outlier detector has to detect the outliers among 12754 points in a 14 dimensional space.

The results show that among these people, H. Lu holds the highest "1st nearest distance", M. Cvetic has the 2nd highest one and K.S. Stelle is the third. Therefore they are chosen as the top three candidates interestingly connecting to Pope.

After analyzing the data, we found that the reason Mr. H.Lu was chosen is that he contributes significantly to a variety of path types. For example, he contributes 35% to the "co-authorship" path with Pope, while the 2nd highest contribution for this type of path from M.Cvetic is only 13%. He also contributes the most (12%) to the "cites paper" path (i.e., Pope writes Paper1, Paper1 cites Paper2, Paper2 is written by H.Lu) and "is cited by path" (Pope writes Paper1, Paper1 is cited by Paper2, Paper2 is written by H.Lu. He also contributes 9% (only surpassed by M. Cvetic's 16%) to the path type "Pope's co-author writes a paper with somebody else" (Pope writes Paper1, Paper1 is also written by Person1, Person1 writes Paper2, Paper2 has another author H.Lu). M. Cvetic and K.S. Stelle are among the top outliers because they also contribute significantly compared with other people except for H. Lu. Figure 2 shows the top 50 outliers (among 12754 others) with their 1st neighbor distance. We can see that there are large gaps between the 1st and 2nd outliers as well as the 2nd and 3rd ones. After that the distance drops rapidly to 0.

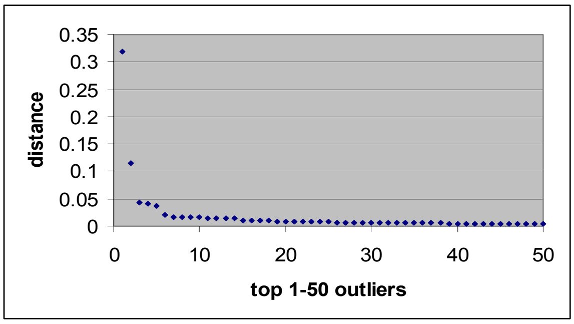

Figure 2 The 1st neighbor distance of the top 50 outliers

The second query we chose was **"which organizations are interestingly connected with Mr. Pope?".** The results show that U. Texas A&M is the most interesting one, followed by SISSA (Scuola Internazionale Superiore di Studi Avanzati) and the third INFN (Italian National Institute for Research in Nuclear and Subnuclear Physics). After analyzing the data we found that the major reason our program regards U. Texas and SISSA to be the outliers is that among the 963 organizations, Pope uses email addresses from only these two institutions. Both institutions contribute 50% in this direction while others contribute 0, which makes them special. However, the reason our program considers INFN as an outlier is different. It is due to the combination of two pieces of evidence:

1. It found that for the majority of institutes the two path types "Pope's colleagues have ever belonged to that institute" and "Pope's co-author belongs to that institute" are positively correlated with respect to the contribution (which also implies that Pope writes many papers with his colleagues in general).

2. It found that the institution INFN contributes the most (8.5%) to the first type shown above, but contributes 0% to the second one (it has no members co-authoring with Pope).

Combing these two facts, our program discovered that INFN is different from others to Pope. In other words, INFN is chosen because Pope tends to write papers with his colleagues but he has never written any with his colleagues that have ever belonged to INFN, despite the fact that INFN produces the most people belonging also to Pope's institution (8.5%). Intuitively this is an interesting and unexpected discovery. There are some possible hypotheses to explain this finding: the first is that Pope's colleagues that are also from INFN might focus only on some specific research direction that does not match with Pope's. The second is that Pope and those who have been to INFN have never been at the same institution at the same time period, thus they did not get a chance to cooperate. The third is that Pope does not get along well with people from INFN; therefore, he does not write paper with them even though they are colleagues.

After investigating through the Wold-Wide Web by combining two interestingly connected nodes as search keywords (e.g. "C.N.Pope SISSA"), we found that Dr. Pope is a professor at U. Texas A&M and he was Dr. H. Lu's thesis advisor (1988-1994). Dr. Cvetic has similar research interests to Pope and works closely with him. Dr. Stelle is a professor of Imperial College London who has ties with Pope not only academically but also personally. For example, Dr. Pope's homepage has a picture showing him and Dr. Stelle traveling together in Afghanistan. Dr. Pope probably was at SISSA, Italy during Fall 1994 and Summer 1996, since his email and mailing address were changed to SISSA during that period. He might also have been there sometime in 1999 as well, since we found from the Web that in his non HEP-Th paper "U duality as general coordinate transformations, and space-time geometry", he used not only the regular Texas A&M address but also an address is SISSA, Italy.

As to the three hypotheses with INFN, we tried to search for the information of Pope's colleagues who also belong to INFN. We found that many of them have similar research interests as Pope, which is a negative justification of the first hypothesis. However, many of them were in SISSA as Pope, but they were not in SISSA with Pope at the same time, which is the positive justification for the second hypothesis. It is not surprising that we did not find any evidence through the Web to support the third hypothesis.

Afterwards we tried to perform several reverse queries. We found that the person that is most interestingly connected with H. Lu is also C.N. Pope. However, when trying to answer "which person is the most interestingly connected to U. Texas A&M", our program indicated that C.N. Pope was only the 3rd outlier, even though academically he contributes much more than others to this institution. The first outlier for this query was Dr. H. Lu, since similar to Pope, many people in this institution either collaborate with Lu or cite or have been cited by his papers. However, it is the fact that Dr. Lu has never belonged to Texas A&M that makes him stand out to become the first outlier (actually, he was there as a student between 1988-1994, but he did not submit any HEP-Th papers using an email address from there).

Next we tried to see how our program performs when using another person as the source. We randomly selected a person, Dr. Chiang-Mei Chen, who has a smaller amount of HEP-Th publications (20) and applied our program to answer the query **"Which organizations are interestingly connected to the person Chiang-Mei Chen?"** The results signify that the school NCU (National Central University, Taiwan) is the 1st outlier followed by MSU (Moscow State University) and NTU

(National Taiwan University). After analyzing the HEP-Th data, we found that they are the only organizations that Dr. Chen has ever belonged to. However, this evidence itself does not make NCU stand out from these three organizations. Looking closer we found that 25% of Chen's co-authors belong to MSU, 14% of them belong to NTU while none of them belong to NCU. As to "citationship", 6% of the papers cited by Chen's paper are from MSU, 2.6% from NTU while none is from NCU. Moreover, 1.6% of the papers citing Chen's paper are from MSU, 0.7% are from NTU and, again, none is from NCU. These facts make MSU and NTU closer to each other from our outlier detector's point of view and, thus, NCU stands out. Intuitively this seems reasonable, since we would expect one would have more co-authorships and citation-ships with the people from the same organization. Thus it is "abnormal" for Chen to have belonged to the school NCU but without any other connection to it. By using the contribution of paths as features, our system first discovered that these three organizations are abnormal to the other 960 ones in the sense that Chen only belongs to them. Furthermore, it finds that NCU is "abnormal among the abnormal", since it is different from the other two. UCSB (5th outlier) is also one organization worthy of noticing, since it contributes the 2nd-most to the papers that cite Chen's work and the 3rd-most to the papers that are cited by Chen's papers. The above facts together with the one that "Chen has never co-authored with any person at that institution" make UCSB a high-ranked outlier.

After checking the Web, we found that Dr. Chen received his Ph.D. from MSU (1999) and then served as an assistant professor at NCU. He has been a postdoc at NTU since August 2002. We found that he is still an assistant professor at NCU, but he kept using NTU's email after 2000.

We also tried to test our program on different types of sources. For example, we chose the keyword "Black hole" as the source for the query **"Which people are interestingly connected to the keyword 'Black Hole'?".** The top four people discovered are Dr. Andrew Strominger, Dr. A.A.Tseytlin, Dr. M.Cvetic and Dr Edward Witten. After manually analyzing the data, we found that the major reason the top three people are outliers is that they wrote a relatively large amount of papers with the keyword "black hole", and they cite or are cited by many other papers that contain "black hole" as keyword. On the other hand, the 4th outlier Edward Witten only published two papers having black hole in the title and did not cite black hole related research frequently. He is abnormal from our outlier detector's point of view because despite not doing research very related to black holes, his papers are still cited by a relatively large amount of papers related to black holes. Moreover, the papers that cite papers with "black hole" as keyword also tend to cite his papers. After investigating the Web through a search engine, we found that the word "black hole" occurred in the research description section of all these four people's homepages. For example, in Dr. Strominger's webpage we found this paragraph "*Strominger and Harvard colleague Vafa gave a statistical derivation of the laws of* ***black hole*** *thermodynamics and the Bekenstein-Hawking entropy formula by counting the quantum microstates of a macroscopic* ***black hole****. This suggests that string theory may yield insight into Hawking's* ***black hole*** *information puzzle…*". We also found that Edward Witten is a famous mathematical physicist who has won the Fields Medal, the highest honor a mathematician can receive. This fact strengthens the validity of our discovery, since even though his research is not mainly on black holes, some of his contributions to the fundamental mathematics are valuable to black hole related research and thus attract many citations.

For the query **"Which season is most interestingly connected to the journal J.Math.Phys?"**, our system returns that Spring 2003 is the first outlier while Winter 2000 is the second. The reason Spring 2003 is the outlier is mainly due to the fact that our program found that no papers in J.Math.Phys cite papers that are published in Spring 2003, which has never happened in any other time period. This makes sense, since the submitting-publishing cycle for this journal could be long and those papers currently in J.Math.Phys might have been submitted before Spring 2003, thus did not have a chance to cite papers published at that time. As to the Winter 2000 period, it is mainly because a relatively large amount of HEP-Th papers published in J.Math.Phys are published during that period, though we could not find an explanation from the Web for this phenomenon.

Analyzing our above experimental results, we find that the novel nodes discovered in the HEP-Th bibliography dataset could be further categorized into two groups: The first group are the nodes that are *significantly* connected to the source and the second are the nodes that are *atypically* connected. In other words, for the bibliography dataset the term "abnormal" can be interpreted as either "significant" or "atypical". For example, H.Lu, U. Texas and Nucl.Phys are significantly connected to Pope and so is Winter 2000 to J.Math.Phys. The reason that the nodes contributing significantly are eminent is that in the HEP-Th bibliography dataset people tend to work with a small number of others, they belong to only a few institutions and usually only focus on a specific research topic. Thus the nodes that are significantly connected with the source become the outliers. On the other hand, our program also detects atypical nodes such as INFN to Pope, NCU and UCSB to Chen, Witten to "black hole" and Spring2003 to J.Math.Phys. These nodes do not contribute the most, but they are picked because they contribute atypically. We also found that in most of the cases we can easily verify the significant nodes through the Web, but not so for the atypical ones. In our opinion this does not mean that the atypical nodes discovered are incorrect, on the contrary, they potentially contain important and interesting information that would be difficult to discover otherwise.

3.3 Novel Loop Discovery

The goal of this experiment is to discover rare loops in the bibliography dataset. We calculated path rarity via 1/N4′ where N4′ is a variation of global fan-out with the additional constraint that source and target have to be the same node. By applying 1/N4' as the rarity measure, our program tries to determine in general which "loop type" is rare in the whole dataset. We measured the rarity of the loops that start from nodes of type "paper". The rarest, least frequent types of loops we found are listed in Table 1.

The rarest loops are papers citing themselves directly, which only occurs 28 times in the whole dataset. We do not have a real world explanation for this and can only attribute it to errors in the dataset. The second, third and fourth paths are citation loops of different length (note that without the loop constraint, these paths are intuitively very common). The rational behind this finding is that for a paper to cite another, the cited paper needs to be published earlier. In this sense a citation loop such as "P1 cites P2 cites P3 cites P1" is really a contradiction in time and should not occur at all. One explanation for such "contradictions" is that sometimes an author (or close colleague) might cite one of his/her own submitted but not yet published papers P2 (which has already cited P1) in a paper P1. The other explanation is that one journal might have a very long revising period and during that period other people can access the previous version. For both explanations we have found supporting instances (e.g. "0110099 cites 0110200 cites 0110186 cites 0110099" for the first case and "9912210 cites 9906151 cites 9509140 cites 9912210" for the second). However, there are still some other unexplainable citation loops (e.g. "9912288 cites 0004011 cites 9911183 cites 9912288") that might occur due to the difference between the true publication date and SLAC-date. The fifth path shows a similar concept where it is rare for a paper to cite another paper that was published during the same time period. This type of loop could also be an indicator for authors that work closely with each other. Finally, the last path shows that people seldom publish multiple papers at the same time. Thus it might be worth further investigation when this type of rare behavior occurs frequently for a particular person.

Top 6 rarest loops
1. PaperX cites PaperX
2. PaperX cites Paper1→Paper1 cites PaperX
3. PaperX cites Paper1→Paper1 cites Paper2 → Paper2 cites PaperX
4. PaperX cites Paper1→Paper1 cites Paper2 → Paper2 cites Paper3 →Paper3 cites PaperX
5. PaperX cites (or cited by) Paper1 → Paper 1 published at Time1→ At Time1, PaperX also published.
6. PaperX is written by Person1 → Person 1 has another Paper1→ Paper1 published at the same time period as PaperX

Table 1: The rare loops

The experiments demonstrate that our approach is capable of uncovering interesting instances masked inside thousands of uninteresting facts. Furthermore, the instances found by novel loop discovery lead us to the discovery of interesting hypotheses or patterns (e.g., that citation loops might be an indicator for authors who work closely with each other or for journals that have a long revision cycle).

3.4 Discussion

The experiments show that our unsupervised program, which has no knowledge about the semantic meanings of paths, can detect interesting connections in an arbitrary network without having to learn rules or patterns. The advantage of our method is that it does everything in an unsupervised manner, thus eliminates the necessity to regenerate new rules or new training data for different queries or even when the whole domain is changed. It also eliminates the risk of being biased by the apparent meanings of the links.

The interesting facts and connections discovered by our program can then focus the user's attention on events that are otherwise hard to be noticed. The insights triggered by such evidences can sometimes lead to the discovery of new patterns or knowledge. For example, without being made aware of those rare loops, we might not ever look into the issue of citation loops at all, since there are thousands of different loops in the dataset that mask this phenomenon. Consequently, we would not discover that citation loops could be an indicator for close authors or journals with long revision periods. The results also prompt the discovery of other related knowledge when one tries to explain them. For example, when we tried to explain why Chen was interestingly connected to NCU, we found that he seldom used that email address.

4. CONCLUSION

We used a notion of "rarity" to measure interestingness in novel path and loop discovery problems and the concept of "abnormality" to model the interestingness of nodes connected to a given source. Our approach is general-purpose and can be applied to any multi-relational dataset that can be represented as a graph of binary relationships. The interestingness and validity of most of our experimental results is either self-evident or is supported by external information found on the Web.

Our approach is different from traditional knowledge discovery. It starts with the automatic identification of potentially interesting instances and connections in the data in an unsupervised manner. Then we manually analyze why those results are considered to be interesting which might lead us to the discovery of interesting hypotheses or patterns. Finally, we try to justify the findings through internal or external sources.

The most important contribution of our methods is that they can discover interesting facts and connections in a dataset in a general, unbiased and unsupervised way without requiring any prior knowledge or training examples for the domain. Potential applications are in homeland security, law enforcement, data cleaning and scientific discovery.

5. ACKNOWLEDGMENTS

Special thanks to Jafar Adibi and Chien-yu Chen for fruitful discussions. This research was sponsored by the Defense Advance Research Projects Agency (DARPA) and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F30602-01-2-0583. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotations thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL or the U.S. Government.

6. REFERENCES

[1] Shou-de Lin and Hans Chalupsky. Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis. in Proceedings of the Third IEEE International Conference on Data Mining. Melbourne, Florida. 2003

[2] R. Rastogi S. Ramaswamy, K. Shim. Efficient Algorithms for Mining Outliers from Large Data Sets. in Proceedings of SIGMOD'00, Dallas, Texas, 2000.

About the Authors

Shou-de Lin is a Ph.D. student in the Computer Science Department of the University of Southern California and a research assistant at USC's Information Sciences Institute.

Dr. Hans Chalupsky is the project leader of the Loom Knowledge Representation & Reasoning Group at USC's Information Sciences Institute.